

Teoría de la Computación y Bases de Datos

Gagliardi, Edilma - Grosso, Alejandro - Herrera, Norma - Reyes, Nora
{oli, agrosso, nherrera, nreyes}@unsl.edu.ar
Proyecto de Investigación "Teoría de la Computación"
dirigido por el Dr. José María Turull Torres
Dpto. de Informática, Universidad Nacional de San Luis

En nuestro trabajo estudiamos las consultas a bases de datos, o queries, usando como marco teórico la Teoría de Modelos Finitos ([AHV95] y [EF95]), con el objetivo de contribuir a la construcción de una Teoría de Bases de Datos que, por una parte, abarque en la forma más completa posible las aplicaciones reales y por otra parte, que admita la consideración de problemas en la computación de queries y en la estructura de las bases de datos que puedan surgir de la utilización de nuevas tecnologías en la materia, como las bases de datos distribuidas.

También buscaremos la caracterización de propiedades de las bases de datos que sean convenientes con relación a la computación de queries con lenguajes incompletos, o a la complejidad de su evaluación, tales como rigidez, L-rigidez, rigidez parcial, L-rigidez parcial y casi rigidez [Tur96], con el fin de estudiar propiedades de bases de datos que permitan aumentar la clase de queries computables en lenguajes incompletos y utilizar estas propiedades relacionadas con la homogeneidad, para una explotación más intensiva y/o eficiente de la información almacenada en las bases de datos.

Ello nos conlleva a estudiar la posibilidad de construir nuevas representaciones o estructuras de datos para clases particulares de bases de datos según las propiedades estudiadas como resultado del punto anterior.

Asimismo hemos pensado en la construcción de bases de datos "cocientes", definidas por un homomorfismo según la relación (de equivalencia) de conmutabilidad por automorfismos en k -tuplas, para obtener así una notable disminución en el tamaño de las bases, preservando no obstante las clases de equivalencia representadas. Esto nos permitiría evaluar los queries sobre las bases cocientes, con mucho menor costo, y trasladar los resultados a las bases originales expandiendo las clases de equivalencia de las k -tuplas.

La Teoría de Modelos, y particularmente la Teoría de Modelos Finitos ([Gur84], [Cai90]) constituye una base teórica sumamente adecuada para el estudio de Bases de Datos, a tal punto que hablar acerca de estructuras relacionales finitas (es decir, en vocabularios que sólo tienen símbolos de relación), es exactamente lo mismo que hablar acerca de instancias de Bases de Datos.

En esta interpretación, para una lógica dada, la noción de expresividad (es decir qué clases son elementales ó EC siguiendo a [Ebi85] y [Flu85]) puede ser considerada como la axiomatizabilidad finita de clases contables de instancias (finitas) de Bases de Datos. Es decir, que una sentencia de una cierta lógica representa un query booleano, y una fórmula, con r variables libres, representa una función de la clase de estructuras finitas de un cierto vocabulario relacional, en la clase de las relaciones r -arias finitas formadas en el dominio de la estructura correspondiente. Es decir, representa un query r -ario.

Hasta el momento se han utilizado mayoritariamente cuatro tipos de formalismos para el estudio de los queries computables: lenguajes de programación formales (QL de [CH80]), máquinas abstractas (GM^{loose} y GM de [AV91]), lógicas abstractas ([EBI85], [Gur84], [Imm87]) y circuitos booleanos [BDG88] [BDG90]. Pero existen diferencias importantes en las clases consideradas en los distintos formalismos. En el ámbito de las máquinas abstractas no se consideran los queries untyped, es decir aquellos para los que la aridad de la relación resultante depende de la base de datos en la que se evalúan. En el ámbito de las lógicas abstractas no se consideran tampoco los queries untyped, ni los queries parciales. Con relación a esta discordancia nos proponemos estudiar las causas que la provocan, e intentaremos, según los resultados, subsanarlas.

Estudiando los modelos de máquinas abstractas, otra meta consistió en construir un modelo de máquina abstracta (en la línea de las máquinas relacionales de [AV91]) que permita computar los queries de aridad variable, o untyped.

Continuando con esta temática, se busca definir una nueva semántica para las lógicas regulares que permita la definibilidad de relaciones cuya aridad dependa de la interpretación. Así, se podría

incorporar la expresión de los queries untyped y estudiar propiedades de los queries, vistos como funciones; permitiendo así definir clasificaciones en la clase de los queries computables distintas a las determinadas por su complejidad computacional. También se podría estudiar la validez de la teoría clásica de complejidad computacional para el caso de los queries untyped y adaptarla o construir una teoría de complejidad ad-hoc, según sea el resultado obtenido, como así también, buscar una formalización para la computabilidad de queries en bases de datos distribuidas.

En el estudio de las lógicas de órdenes superiores (mas allá de Primer Orden) en cuanto a la expresividad de queries computables, hemos considerado aplicaciones de interés y analizado las limitaciones relevantes. Entonces, analizamos consultas a bases de datos relacionales expresadas utilizando fórmulas de extensiones de lógica de primer orden con distintos tipos de cuantificadores, que suponen diferentes tipos de iteraciones en su interpretación semántica [AVV97][EF95]. Con FO se captura sobre estructuras finitas ordenadas una clase estrictamente contenida en DSPACE(log n) y con estas extensiones de FO se captura la clase EXPTIME.

En el estudio de la expresividad de diferentes lógicas por debajo de segundo orden (SO) la presencia de una relación de orden total en las estructuras (es decir, un símbolo de relación binario en sus vocabularios siempre interpretado por una relación de orden total en el dominio de la estructura dada) surgió como algo de gran relevancia ([Imm82], [Gur88], [DM89], [BIS90], [IPS91], [AV91]) a tal punto que se tornó usual el hablar acerca de lenguajes con una relación de orden total (así asumida) en sus vocabularios. Por ejemplo, se sabe que Primer Orden (FO) es menos expresivo que (FO+FP) (esto es, FO enriquecido con un cuantificador de Punto Fijo que permite iteraciones que siempre terminan) porque en FO no podemos expresar propiedades de clausura tales como la existencia de un ciclo en un grafo ([AU79]); el que a su vez es menos expresivo que (FO + FP + \leq) (la misma lógica, pero sólo aplicada a estructuras que se asumen ordenadas). Este último es equivalente a la clase P de las estructuras decidibles en tiempo polinomial en el modelo de la Máquina de Turing ([Imm87]). En esta clase podemos expresar la paridad del dominio de la estructura de entrada, no siendo así sin \leq ([CH82]).

La computación de queries o consultas a Bases de Datos muy grandes o distribuidas requiere de un marco teórico sólido que provea herramientas formales para su consideración. Véanse al respecto los problemas descritos en [Tur95] con relación a la inadecuación de la Teoría de Computabilidad clásica con respecto a la computación de queries a Bases de Datos. Esto provoca problemas prácticos dado que los motores de Bases de Datos permiten expresar queries que no preservan isomorfismos (como los datos de "un" cliente que cumpla cierta propiedad). Si se desea parcializar la evaluación de tal query (por ejemplo porque su evaluación requiere mucho tiempo de procesador) y realizar una reorganización de la Base de Datos entre dos computaciones parciales, se pueden lograr resultados inconsistentes, y esto es debido a que dichos motores no respetan el concepto de query computable.

El punto esencial es que en la medida en que la Informática pretende atacar problemas de alta complejidad (lo que se ha tornado habitual debido al tremendo desarrollo en cuanto a tecnología informática) se hace más necesaria una teoría de base, sólida y dinámica, que permita utilizar herramientas formales para enfrentar esos problemas. De allí que pensemos que es altamente relevante hoy en día el desarrollo de las teorías de base que permitan comprender cabalmente los fenómenos que subyacen en las distintas áreas de las Ciencias de la Computación, fenómenos que utilizamos, pero que no alcanzamos a comprender plenamente.

Por otra parte, estos desarrollos teóricos, nos permitirán definir pautas de diseño en las Bases de Datos que apunten a evitar los posibles conflictos que se desprenden de la inadecuación de los motores de Bases de Datos existentes a la teoría, antes mencionada.

En el marco de las familias de Circuitos Booleanos se pueden considerar aplicaciones donde la cardinalidad del dominio sea constante y tal que los queries expresados como una subfamilia de circuitos booleanos, constituyen un marco apropiado para estudiar el grado de paralelización de la consulta.

En cuanto a optimización de queries y lenguajes de aplicación, se deben considerar estrategias de procesamiento que minimicen el uso de los recursos. Para ello, se han analizado estrategias de optimización estáticas, tal que tomando un query expresado en cálculo relacional se obtiene una expresión en álgebra relacional optimizada para su evaluación.

Se pretende consolidar y alimentar esta línea de trabajo con la supervisión de Trabajos Finales de la Licenciatura en Ciencias de la Computación que se vinculen con el tema, y también desarrollar

seminarios, y eventualmente materias optativas al nivel de grado y de post-gradó para la carrera y posiblemente para otras carreras; dado que se involucran temas como el de Lógica, Teoría de Modelos, Teoría de Computabilidad y Teoría de Complejidad. Dichas temáticas pueden ser de también de interés para las carreras de Matemáticas. En 1994 se protocolizó el Seminario de Teoría de la Computación dirigido por el director del proyecto, y que se estaba llevando a cabo desde 1992. También el director del grupo presentó un plan de Maestría en Ciencias de la Computación con orientación en Teoría de la Computación que viene funcionando desde 1996. Dicha maestría lleva 4 años de iniciada, contando con dos promociones, encontrándose varios de ellos en la etapa de definición de sus respectivas tesis.

REFERENCIAS

- [AHV95] Abiteboul, S.; Hull, R.; Vianu, V.; "Foundations of Databases", Addison-Wesley Publishing Company, 1995.
- [AU79] Aho, A.V.; Ullman, J.D.; "Universality of Data Retrieval Languages", in Proceedings 6th ACM Symp. on Principles of Programming Languages, 1979.
- [AV91] Abiteboul, S.; Vianu, V.; "Generic Computation and Its Complexity", STOC 1991.
- [AVV97] Abiteboul, S.; Vardi, M. y Vianu, V. "Fixpoint Logic, Relational Machines and Computational Complexity", Journal of ACM, 44(1), 30-56. 1997.
- [BDG88] Balcázar, J.L.; Díaz, J.; Gabarró, J.; "Structural Complexity I". Springer- Verlag, 1988.
- [BDG90] Balcázar, J.L.; Díaz, J.; Gabarró, J.; "Structural Complexity II". Springer- Verlag, 1990.
- [BIS90] Barrington, D.; Immerman, N.; Straubing, H.; "On Uniformity within NC", Journal of Computer and System Sciences 41, 1990.
- [Cai90] Caicedo, X.; "Elementos de Lógica y Calculabilidad", Empresa Docente, Universidad de los Andes, 1990.
- [CH80] Chandra, A.; Harel, D.; "Computable Queries for Relational Data Bases", Journal of Computer and System Sciences 21, 1980.
- [CH82] Chandra, A.; Harel, D.; "Structure and Complexity of Relational Queries", Journal of Computer and System Sciences 25, 1982.
- [DM89] Dublisch, P.; Maheshwari, S.; "Expressibility of Bounded Arity Fixed-point Query Hierarchies", PODS 1989.
- [Ebi85] Ebbinghaus, H.; "Extended Logics: The General Framework", in Perspectives in Mathematical Logic, Springer-Verlag, 1985.
- [EF95] Ebbinghaus, H.; Flum, J.; "Finite Model Theory", Springer-Verlag, 1995.
- [Flu85] Flum, J.; "Characterizing Logics", in Perspectives in Mathematical Logic, Springer-Verlag, 1985.
- [Gur84] Gurevich, Y.; "Toward logic tailored for computational complexity", in Computation and Proof Theory, Lecture Notes in Mathematics 1104, 1984.
- [Gur88] Gurevich, Y.; "Logic and the Challenge of Computer Science", in Trends in Theoretical Computer Science, Computer Science Press, 1988.
- [Imm82] Immerman, N.; "Upper and Lower Bounds for First Order Expressibility", Journal of Computer and System Sciences 25, 1982.
- [Imm87] Immerman, N.; "Languages that Capture Complexity Classes", SIAM J. on Computing 16, 4; 1987.
- [IPS91] Immerman, N.; Patnaik, S.; Stemple, D.; "The Expressiveness of a Family of Finite Set Languages", PODS 1991.
- [Tur95] Turull Torres, J.M.; "De la Computabilidad sobre Números Naturales a la Computabilidad sobre Estructuras o Bases de Datos", INFOCOM 95, Junio 1995.
- [Tur96] Turull Torres, José María; "Clases de Bases de Datos L-rígidas y Expresividad de Lenguajes Relacionales Incompletos". Tesis Doctoral, Universidad Nacional de San Luis, 1996.